MASTER'S THESIS

Invloed van Formatieve Oefentoets op Hogere Orde Kennis Eindtoets

Szczerba, Petra

Award date:
2021

Open Universiteit

Invloed van Formatieve Oefentoets op Hogere Orde Kennis Eindtoets

Influence of Formative Practice Test on Higher Order Knowledge Final Test

Petra Szczerba

Master Onderwijswetenschappen
Open Universiteit

Cursusnaam en cursuscode:	OM9906 - Masterthesis
Naam begeleider:	Prof. Dr. D. Joosten – Ten Brinke
Datum:	3 februari 2021

Voorwoord

The harder you work for something, the greater you'll feel when you finally achieve it.

Het is klaar. Na ruim drie jaar studie ligt mijn masterthesis voor u. Met erg veel plezier en af en toe met tegenzin heb ik dit sluitstuk geschreven. Mijn ervaring is dat dit werkstuk een mooie samenvatting is van alle eerdere premaster- en mastermodules die gevolgd zijn. Alle kennis die daarin is opgedaan, komt samen. Ondanks tussentijdse hobbels die soms onmogelijk te overkomen leken, met name de coronapandemie, kijk ik terug op een leerzame periode waarin ik naast opgedane kennis ook veel gezellige en inspirerende medestudenten heb leren kennen, in het bijzonder Ann, Gemma, Meike en Caroline. Zij hebben een waardevolle bijdrage geleverd aan mijn studiesucces en -plezier. Goede vriend Marnix wil ik bedanken voor zijn oeverloze geduld en goede uitleg als we aan het sparren waren over mogelijke data-analyses. Mede door hem ben ik statistiek interessanter gaan vinden. Ik wil mijn begeleidster Desirée Joosten-Ten Brinke bedanken voor de steun en het optimisme dat ze in alle overlegmomenten tentoonspreidde. Dit gaf mij telkens weer de moed om door te gaan en het proces van de zonnige kant te blijven zien. Examinator José Janssen bedank ik voor de opbouwende feedback tijdens het schrijfproces. Dit heeft het stuk goed gedaan! Vanzelfsprekend ben ik mijn collega's en studenten van de opleiding verpleegkunde waar ik werkzaam ben erg dankbaar dat zij bereid zijn geweest tijd te stoppen in het ondersteunen van en het meewerken aan mijn onderzoek. Zonder hen had ik deze thesis niet kunnen afronden. Ten slotte wil ik mijn gezin bedanken dat ze mij ontlastten tijdens piekperiodes waardoor ik op de juiste momenten de focus daar kon leggen waar het nodig was. Vanaf nu zijn de weekenden en avonden weer voor jullie!

Petra Szczerba
Eindhoven, februari 2021

Inhoud

Samenvatting

Formatief toetsen is ondersteunend aan het leerproces en draagt bij aan de ontwikkeling van kennis en vaardigheden. Over de voorwaarden waaronder formatief toetsen effectief is, is weinig onderzoek verricht. De geheugenpsychologie heeft echter wel veel onderzoek gedaan naar effectieve leerstrategieën die overeenkomsten vertonen met processen van (tussentijds) formatief toetsen (Adesope, Trevisan, & Sundararajan, 2017; Roediger & Karpicke, 2006). Vanuit een overzichtsstudie die verricht is door Dirkx, Joosten-Ten Brinke, en Camp (2019) is een tiental richtlijnen voor formatieve toetsing ontwikkeld. Aanvullende feedback kan hierbij waardevol zijn (Roediger & Butler, 2011; Rowland, 2014). De inzichten van waaruit deze richtlijnen tot stand zijn gekomen zijn, zijn afkomstig van onderzoeken gericht op kennis en vaardigheden van lagere orde. Er is tot op heden weinig onderzoek gedaan naar de effecten van effectieve leerstrategieën bij hogere orde kennis en vaardigheden. Van studenten in het hoger onderwijs worden deze echter wel verwacht.

Het doel van dit onderzoek was om na te gaan wat het effect was van een oefentoets, rekening houdend met bovengenoemde richtlijnen, op de resultaten van een hogere orde kennis eindtoets. Daarbij is aanvullend de rol van feedback bij formatieve toetsing meegenomen.

Een experimenteel onderzoek is uitgezet onder alle eerstejaarsstudenten van een verpleegkundeopleiding aan een onderwijsinstelling voor hoger onderwijs. Alle propedeusestudenten (n = 552) zijn uitgenodigd om deel te nemen aan dit onderzoek waarvan 435 studenten meededen. De controlegroep bestond uit studenten van een eerder onderwijscohort (n = 473). Deelnemers aan het onderzoek ontvingen een oefentoets in lesweek 5, halverwege de onderwijsmodule, waarbij ad random de ene helft van de deelnemers juist/onjuist feedback ontvingen en de andere studenten dezelfde oefentoets kregen met toelichtende feedback. De data zijn geanalyseerd met behulp van ANOVA, ANCOVA, Pearson's correlatie en regressieanalyse. Uit de resultaten van het onderzoek bleek dat studenten die een oefentoets hebben gemaakt, niet significant hoger scoorden op de eindtoets. Dit gold voor beide experimentele groepen. Ook het type feedback liet geen verschil zien in de eindscores. Voor de scores op de oefentoets was wel een significant effect zichtbaar. De hoogte van de scores op de oefentoets lieten een beperkte samenhang zien met de hoogte van de scores op de eindtoets. De resultaten van het voorliggende onderzoek tonen niet aan dat een oefentoets, ongeacht de vorm van de feedback, effectief is voor de resultaten op de hogere orde kenniseindtoets. Een aanbeveling is om bij soortgelijk vervolgonderzoek op zoek te gaan naar toetsen waarvan het leerdoel betrekking heeft op analyseren, evalueren of creëren. Deze worden in de wetenschappelijke literatuur als hogere orde kennis en vaardigheden beschouwd in tegenstelling tot het toepassen van kennis. Over deze laatste zijn de opvattingen verdeeld.

Trefwoorden: formatieve toets, feedback, hogere orde kennis, retrieval practice

Summary

Formative assessment is supportive of the learning process and contributes to the development of knowledge and skills. Little research has been done on the conditions under which formative assessment is effective. However, memory psychology has conducted a lot of research on effective learning strategies; these have similarities with the processes of (interim) formative testing (Adesope, Trevisan, & Sundararajan, 2017; Roediger & Karpicke, 2006). From a review study conducted by Dirkx, Joosten-Ten Brinke, and Camp (2019), a dozen guidelines for formative assessment were developed. Additional feedback can be valuable here (Roediger & Butler, 2011; Rowland, 2014). The insights from which these guidelines were created have come from studies focused on lower-order knowledge and skills. To date, little research has been done on the effects of effective learning strategies with respect to higher order knowledge and skills. However, these are expected of students in higher education. The purpose of this study was to investigate the effect of a practice test, taking into account the above guidelines, on the results of a higher order knowledge final test. This additionally included the role of feedback in formative testing.

An experimental study was carried out among all first-year nursing students at an educational institution of higher education. All first-year students (n = 552) were invited to participate in this study of which 435 students participated. The control group consisted of students from a previous educational cohort (n = 473). Participants in the study received a practice test in lesson week 5, halfway through the teaching module, where ad random one half of the participants received correct/incorrect feedback and the other students received the same practice test with explanatory feedback. The data were analyzed using ANOVA, ANCOVA, Pearson's correlation and regression analysis. The results of the study showed that students who took a practice test did not score significantly higher on the final test. This was true for both experimental groups. The type of feedback also produced no difference in the final scores. A significant effect was visible for the scores on the practice test, however. The height of the scores on the interim test showed a limited correlation with the height of the scores on the final test. The results of the present study do not show that a practice test, regardless of the form of the feedback, is effective for the results on the higher order knowledge final test. A recommendation is that similar follow-up research should consider tests where learning objectives relate to analyzing, evaluating, or creating. These are considered higher order knowledge and skills in the scientific literature as opposed to the application of knowledge. Views on this are divided.

Keywords: formative assessment, feedback, higher order knowledge, retrieval practice

1. Inleiding

Formatieve toetsen dragen bij aan het ontwikkelen van kennis en vaardigheden. Het biedt het leerproces ondersteuning, wat vervolgens kan leiden tot betere leeropbrengsten. Het is dan wel noodzakelijk om deze toetsen op de juiste manier in te zetten (Van Berkel, Bax, & Joosten-Ten Brinke, 2017). Over de voorwaarden waaronder formatieve toetsen een positieve invloed hebben op leeropbrengsten is echter weinig bekend. Wel hebben wetenschappers uit de geheugenpsychologie onderzoek gedaan naar effectieve leerstrategieën, die overeenkomsten vertonen met de processen van formatief toetsen, zoals retrieval practice (oefenen met het ophalen van informatie uit het geheugen) en distributed practice (gespreid oefenen) (Adesope, Trevisan, & Sundararajan, 2017; Roediger & Karpicke, 2006). Om de resultaten van onderzoeken vanuit de geheugenpsychologie te koppelen aan kennis over het proces van (tussentijds) formatief toetsen is onlangs een overzichtsstudie uitgevoerd (Dirkx, Joosten-Ten Brinke, & Camp, 2019). Dit heeft geresulteerd in tien ontwerprichtlijnen voor formatieve toetsen (Dirkx et al., 2019; zie Tabel 1).

Tabel 1

Tien ontwerprichtlijnen voor gebruik van formatieve toetsen

	Beschrijving
1	Gebruik formatieve toetsen in verschillende domeinen en bij verschillende soorten leermaterialen om leren te stimuleren;
2	Gebruik formatieve toetsen in elk geval voor onthouden, begrijpen, en toepassen van informatie;
3	Stem het niveau en de inhoud van de formatieve toets af op de eindtoets;
4	Kies voor een combinatie van open- en gesloten vragen bij formatieve toetsen;
5	Als je formatief toetst, zorg dan dat je in de feedback het goede antwoord geeft;
6	Zet een formatieve toets pas in na een initiële leerfase;
7	Toets dezelfde stof minstens één keer maar maximaal drie keer;
8	Spreid de toetsen uit over de tijd;
9	Begin niet vlak voor de summatieve toets met het maken van formatieve toetsen maar gebruik 20% regel;
10	Bed formatieve toetsen bewust in het toetsprogramma in, waarbij de programmering geen vrijblijvend maar sturend karakter heeft.

De overzichtsstudie laat zien dat de resultaten uit onderzoeken van waaruit deze richtlijnen voortvloeien robuust zijn bij lagere orde cognitieve kennis en vaardigheden (onthouden en begrijpen in vergelijkbare opgaven; zie richtlijn 2). Of de richtlijnen een vergelijkbaar effect hebben op hogere orde kennis en vaardigheden, is op dit moment te weinig onderzocht om daar eenduidige uitspraken over te

doen. In het hoger onderwijs is echter juist steeds meer aandacht voor deze hogere orde vaardigheden. Daarom is het doel van dit onderzoek om na te gaan of de ontwikkelde richtlijnen voor de lagere orde kennis ook op hogere orde kennis van toepassing zijn.

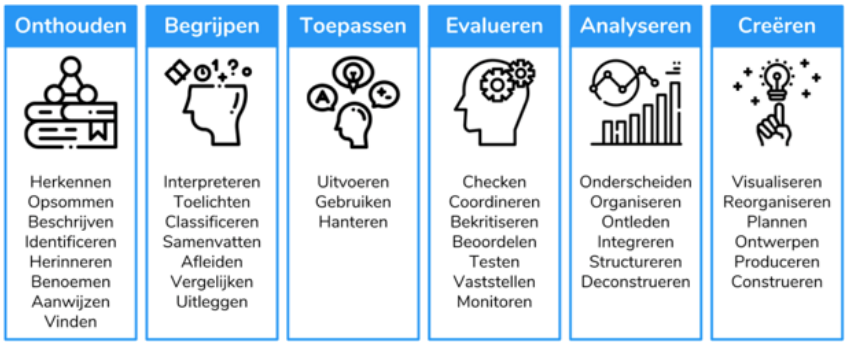
1.1 Theoretische kader

1.1.1 Hogere orde kennis en vaardigheden

De ontwikkeling van hogere orde denkvaardigheden is essentieel binnen het onderwijs om leren te bewerkstelligen. Er zijn diverse definities voorhanden die hogere orde kennis en vaardigheden beschrijven. Brookhart (2010) identificeert hierin drie categorieën, te weten 'transfer', 'kritisch denken' en 'probleemoplossing'. Transfer houdt in dat studenten niet slechts onthouden maar ook begrijpen wat ze hebben geleerd en hoe ze het kunnen gebruiken in nieuwe situaties (Anderson & Krathwohl, 2001). Kritisch denken houdt in dat je in staat bent om reflectief te denken waarbij je zelfstandig komt tot weloverwogen oordelen, afwegingen en beslissingen. Het is een gestructureerde manier om ervaringen te begrijpen, te analyseren en er betekenis aan te geven. Vaardigheden die vallen onder kritisch denken zijn redeneren, vragen stellen, onderzoeken, observeren, beschrijven, vergelijken en verbinden en het bepalen van diverse gezichtspunten (Barahal, 2008). Probleemoplossing kan gedefinieerd worden als: "een student heeft een probleem wanneer hij een specifiek resultaat of doel wil bereiken, maar niet automatisch het juiste pad of de juiste oplossing herkent om dit te bereiken. Om het gewenste doel te bereiken, moet hij probleemoplossend denken" (Nitko & Brookhart, 2007, p. 4). Algemener geformuleerd is probleemoplossing een vaardigheid die een persoon in staat stelt om een oplossing te vinden voor een probleem dat niet opgelost kan worden op basis van wat eerder onthouden is. Deze problemen hebben een open einde en kunnen meerdere oplossingen bevatten. Bovenstaande definities zijn gerelateerd aan de taxonomie van Bloom, waarin een onderscheid wordt gemaakt in de mate van complexiteit van vaardigheden. Bij het ontstaan van de taxonomie leek deze hiërarchisch georganiseerd. Men ging ervan uit dat een hoger niveau in de taxonomie de lagere niveaus veronderstelt (Van Berkel et al., 2017). In 2001 is de taxonomie aangepast en betreft het met name een beschrijving van verschillende soorten leeractiviteiten die niet per se hiërarchisch van aard zijn. In deze versie vallen 'onthouden, 'begrijpen', en 'toepassen' onder de minder complexe 'lagere orde' vaardigheden, en 'analyseren', 'evalueren' en 'creëren' onder de 'hogere orde' vaardigheden (Anderson & Krathwohl, 2001). In Figuur 1 wordt de taxonomie van Bloom weergegeven waarin de hiërarchie ontbreekt (Lucassen, 2018). Uit bovenstaande blijkt dat de genoemde auteurs verschillende (sub)categorieën van hogere denkvaardigheden hanteren. Bijvoorbeeld in het geval van 'toepassen van kennis' scharen Anderson en Krathwohl (2001) dit onder de lagere orde kennis en vaardigheden terwijl Brookhart (2010) dit bij hogere orde denkvaardigheden indeelt.

Recent onderzoek naar complexe cognitieve processen laat zien dat het ophalen van informatie uit het geheugen een effectieve strategie is om de transfer van leren van studenten te bevorderen. In het

onderzoek van Agarwal (2018), dat de taxonomie van Bloom als uitgangspunt neemt, wordt aangetoond dat een oefentoets bij een hogere-orde-eindtoets pas effectief is als de vragen op de oefentoets op hetzelfde niveau gesteld moeten worden. Het verbetert het leren van hogere orde vaardigheden en taken. Het oefenen met behulp van bijvoorbeeld een feitentoets heeft weinig effect op een toets waarop hogere orde kennis vereist is (Agarwal, 2018). Het aantal uitgevoerde onderzoeken naar oefentoetsen voor hogere orde kennis en vaardigheden is echter beperkt.



Figuur 1. Herziene taxonomie van Bloom (Lucassen, 2018).
Noot. Herdrukt van "Taxonomie van Bloom,", door Lucassen, M. 2018, 14 november. Geraadpleegd van https://www.vernieuwenderwijs.nl/de-taxonomie-van-bloom-vaak verkeerd-gebruikt-maar-zo-werkt-het-wel/

1.1.2 Formatieve toetsing

De onderwijsraad heeft geconstateerd dat de huidige toetspraktijk van het Nederlandse onderwijs niet of onvoldoende bijdraagt aan de onderwijskwaliteit. Bij de huidige toetsing ligt de nadruk op de summatieve functie van toetsing waarbij het gaat om selectie en is er te weinig ruimte voor de formatieve functie. Door de grote hoeveelheid beslissende toetsen ervaren studenten de studie als hordeloop. (Onderwijsraad, 2018). Formatieve toetsing heeft als doel het leerproces te verbeteren en inzicht te krijgen en geven in het leerproces. Dit inzicht is nodig om het leerproces te sturen. Een eenduidige formulering van het begrip 'formatief toetsen' is echter niet voorhanden (Sluijsmans, Joosten-Ten Brinke, & Van der Vleuten, 2013). De ontwikkeling van dit begrip is door Brookhart (2007) beschreven. Formatief toetsen werd rond de jaren zeventig vooral ingezet voor informatie met betrekking tot het leerproces (Scriven, 1967). Al snel werd dit uitgebreid omdat het ook nuttige informatie voor de docenten kan opleveren in het kader van onderwijsbeslissingen, waarna ook het belang voor de studenten werd toegevoegd in de zin dat zij zelf deze formatieve toetsing kunnen inzetten voor hun eigen leren. Tegenwoordig is het doel van formatieve toetsing dat deze, naast het leerproces, ook bijdraagt aan de motivatie van studenten (Brookhart, 2007).

Formatief toetsen kan op een informele of formele manier plaatsvinden. Bij de informele aanpak worden de formatieve toetsen geïntegreerd in de dagelijkse onderwijspraktijk. Van formele toetsing is

sprake als er met genormeerde toetsen op vastgestelde momenten wordt gewerkt (Sluijsmans et al., 2013). De feedback bij deze genormeerde toetsen maakt de toetsen formatief. In de literatuur worden drie benaderingen van formatief toetsen onderscheiden: opbrengstgericht werken, assessment for learning en diagnostische toetsen (Schildkamp et al., 2014). Bij opbrengstgericht werken worden op systematische wijze gegevens verzameld en geanalyseerd. Deze analyses leveren input op om het onderwijs te kunnen verbeteren. Assessment for learning heeft betrekking op het richting geven aan het leerproces van leerlingen: "…het proces van zoeken, aggregeren, interpreteren van informatie die studenten en docenten gebruiken om te bepalen waar studenten staan in hun leerproces, waar zij naar toe moeten en op welke manier" (Assessment Reform Group, 2002, p. 1). Om de ontwikkelingsfase en de daarbij behorende leerbehoeften van individuele studenten te bepalen worden diagnostische toetsen ingezet (Schildkamp et al., 2014). Voor de instrumentele definitie zijn meerdere benamingen voorhanden, zoals 'oefentoets', 'diagnostische toets', 'quiz', 'deeltoets' of 'tussentijdse toets'. In dit onderzoek is gekozen om de term 'oefentoets' te hanteren.

1.1.3 Effecten van oefentoetsen

Veel studies laten zien dat oefentoetsen het leren van studenten bevorderen (Black & Wiliam, 1998; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Deze toetsen kunnen meerdere effecten ten aanzien van het leren van studenten teweegbrengen. De eerste betreft de pre-effecten van leren. Hierbij gaat het om de manier waarop de studenten zich voorbereiden op een toets, op basis van de verwachtingen over de toets. Post-effecten gaan over de wijze waarop bijvoorbeeld feedback na het maken van de toets bijdraagt aan leren. Pure-effecten van toetsing verwijzen naar de effecten die het maken van een toets heeft op de leerprestaties (Dochy, Segers, Gijbels, & Struyven, 2007). Het laatstgenoemde effect wordt vooral bestudeerd vanuit de geheugenpsychologie (Adesope et al., 2017; Roediger & Karpicke, 2006).

Een begrip dat vanuit de geheugenpsychologie gerelateerd is aan oefentoetsen is testing effect. Dit fenomeen is al meer dan een eeuw geleden onderzocht (Foss & Pirozzolo, 2017). Het verwijst naar leerwinst en het onthouden dat kan optreden wanneer studenten een oefentoets maken over materiaal dat ze moeten kennen voor een eindtoets. De definitie van het testing effect is: "de positieve invloed van oefentoetsen op toekomstig leren en onthouden voor de eindtoets in vergelijking met veel gebruikte leerstrategieën zoals herlezen of concept-mapping" (Adesope et al., 2017, p. 660). Er zijn twee verklaringen voor de positieve invloed van oefentoetsen op het leren. De eerste is de *Transfer Appropriate Processing* hypothese (TAP). De geheugenprestatie hangt af van de overlap tussen de coderings- en ophaalprocessen uit het geheugen. Deze hypothese gaat ervan uit dat het ophalen de geheugenprestaties verbetert als gevolg van vergelijkbare processen die voor zowel de oefentoets als de eindtoets nodig zijn. De andere verklaring is de *Retrieval Effort* hypothese (RE).

Deze gaat ervan uit dat hoe moeilijker het de lerende gemaakt wordt om de informatie uit het geheugen op te halen tijdens de oefentest, des te beter hetzelfde leermateriaal wordt onthouden (Butler, 2010; Roediger & Butler, 2011; Stenlund, Sundström, & Jonsson, 2014).

De mogelijke effecten van retrieval practice op het leren en onthouden van informatie is de afgelopen 15 jaar door de geheugenpsychologie bestudeerd (Adesope et al., 2017). De letterlijke vertaling van retrieval practice is: "oefenen met het ophalen van informatie uit het geheugen*"* (Dirkx et al., 2019, p. 9). Dit ophalen uit het geheugen vindt veelal plaats met behulp van oefentoetsen zoals besproken in paragraaf 1.1.2. In diverse overzichtsstudies is aangetoond dat retrieval practice helpt bij het onthouden van nieuwe kennis (Dunlosky et al., 2013; Karpicke & Aue, 2015, Moreira, Pinto, Starling, & Jaeger, 2019; Roediger & Karpicke, 2006; Roediger & Pyc, 2012). Het beter scoren op de eindtoets door retrieval practice wordt als het *directe* effect beschouwd. Er zijn ook *indirecte* effecten te benoemen, zoals een betere inschatting van het eigen leren, hogere motivatie en betere sturing van het eigen leerproces (Roediger & Karpicke, 2006). In de meeste onderzoeken wordt het effect van retrieval practice vergeleken met minder effectieve leerstrategieën zoals het opnieuw bestuderen van de leerstof. Resultaten laten zien dat het lange termijn onthouden in het geval van retrieval practice significant beter is bij studenten die oefenen met het ophalen van informatie uit het geheugen met behulp van een oefentoets dan bij de groep die de stof nogmaals bestudeert. Dit geldt zowel voor onderzoek dat uitgevoerd is in het laboratorium als onderzoek in de onderwijspraktijk, bij verschillende leermaterialen (woordenlijsten, leerboeken, topografie), verschillende toetsvragen (o.a. multiple choice vragen, invulvragen, kort-antwoord) en verschillende doelgroepen (van basisschoolleerlingen tot volwassenen) (Dirkx et al., 2019). Van Gog en Sweller (2015) benoemen twee mogelijke verklaringen waarom retrieval practice werkt bij het leren. De eerste is dat de informatie georganiseerd wordt tijdens retrieval practice, waardoor bruikbare associaties tussen de informatie-eenheden worden gelegd. Het bewustzijn van deze relaties tussen begrippen draagt bij aan betere prestaties op de eindtoets. Het ophalen van begrippen wordt gemakkelijker door de vindbaarheid van andere items waaraan deze begrippen zijn gekoppeld. Als tweede verklaring wordt de versterking van de geheugensporen genoemd (Van Gog & Sweller, 2015). Deze kunnen zowel direct via retrieval practice of indirect via feedback worden versterkt.

1.1.4 Feedback

Uit de onderzoeken van onder andere Roediger en Butler (2011) en Rowland (2014) blijkt dat aanvullende feedback, gegeven door een docent, bij een oefentoets positieve effecten kan bewerkstelligen voor de scores op eindtoetsen. In het bijzonder als de beheersing op de items van een oefentoets beperkt is. Een definitie van goede feedback, volgens Nicol en Macfarlane-Dick (2006) is: "feedback die studenten helpt hun eigen prestaties te verbeteren en te corrigeren, het helpt studenten actie te ondernemen om de verschillen tussen hun intentie en behaalde resultaten te verkleinen" (Nicol & Macfarlane-Dick, 2006, p. 208). Black en Wiliam (1998) onderscheiden vier niveaus van feedback,

te weten; feedback op taakniveau (in hoeverre wordt de taak goed uitgevoerd), procesniveau (hoe verloopt het proces), zelfregulatieniveau (het zelf reguleren van acties en het eigen gedrag monitoren) en zelf niveau (feedback op de persoon en niet op de taak). Uit de meta-analyse van Hattie en Timperley (2007) blijkt dat voornamelijk feedback op procesniveau en zelfregulatieniveau effectief is voor de leeropbrengsten.

Glover en Brown (2006) hebben onderzocht, ongeacht het niveau waarop de feedback wordt gegeven, op welke verschillende manieren feedback het meest effectief is. Ten eerste dient de feedback regelmatig gegeven te worden. Hierbij is het van belang dat er niet te veel tijd tussen de toets zit en de feedback die erop volgt. Ten tweede moet de feedback, gekoppeld aan het doel van de taak die de studenten moeten uitvoeren, begrijpelijk geformuleerd zijn. Ten slotte moet het gericht zijn op leren (feedforward; Glover & Brown, 2006). De studie van Glover en Brown (2006) heeft hiernaast ook een indeling gemaakt met betrekking tot de diepgang van feedback. Deze is in drie categorieën te verdelen (zie Tabel 2).

Tabel 2

Categorieën diepgang van feedback

	Beschrijving	Ter illustratie ‘spelfout’
Categorie 1	Een probleem wordt gesignaleerd. Er wordt een zwakte vastgesteld zonder advies hoe het beter kan of moet.	Een spelfout wordt aangestreept met een rode pen.
Categorie 2	Een probleem wordt gesignaleerd en er wordt tevens een correctief advies gegeven of er wordt verwezen naar bronnen waar het goede antwoord te vinden is.	Een spelfout wordt gecorrigeerd en/of er wordt verwezen naar de spellingregels in het lesboek.
Categorie 3	Een probleem wordt gesignaleerd en er wordt een correctief advies. Daarnaast wordt een verklaring gegeven voor de fout en/of de aard van de correctie. Er zit ook een element van feed forward in.	Een spelfout wordt gecorrigeerd en/of verwezen naar de spellingregels. Daarbij wordt uitleg gegeven waarom het gegeven antwoord niet correct is met een advies waar in de toekomst op gelet moet worden bij deze spellingsregel.

In het onderzoek naar testing effect wordt zowel categorie 1 als categorie 2 feedback gegeven. Glover en Brown (2006) doen geen uitspraak over welke categorie het meest effectief is. Vanuit de definitie van Nicol en Macfarlane-Dick (2006) is echter te verwachten dat categorie 2 feedback nuttiger is dan categorie 1.

1.1.5 Retrieval practice bij hogere orde vaardigheden

Van Gog en Sweller (2015) wijzen erop dat onderzoeksresultaten met betrekking tot de effectiviteit van retrieval practice, bij hogere orde kennis en vaardigheden, genuanceerd moeten worden. Zij halen meerdere onderzoeken aan waaruit blijkt dat wanneer het beroep op hogere orde kennis en vaardigheden toeneemt, het effect van retrieval practice (testing effect) afneemt. Door de sterke verbindingen tussen de verschillende elementen bij complexe leertaken, valt het voordeel van retrieval practice deels weg. De associaties zijn namelijk reeds voorhanden. De organisatorische verwerking die plaatsvindt door het testen, heeft geen toegevoegde waarde. Echter, de bevindingen van Van Gog en Sweller (2015) worden weerlegd door Karpicke en Aue (2015). Zij zijn het oneens met de stelling dat het effect van retrieval practice nagenoeg verdwenen is bij complexere leertaken die hogere orde denkvaardigheden vereisen. Met name methodologische oorzaken in het onderzoek van Van Gog en Sweller (2015) worden hiervoor door hen genoemd; de literatuurstudie zou niet consistent uitgevoerd zijn. Onderzoeken die wel hebben laten zien dat retrieval practice van nut is bij complexe leermaterialen zijn buiten beschouwing gelaten. Een voorbeeld van een onderzoek dat de bevindingen van Van Gog en Sweller tegenspreekt, is van Dobson, Linderholm, en Perez (2018). Studenten moesten in dit onderzoek complexe fysiologische informatie evalueren. Bij herhaald studeren scoorden studenten significant lager op de eindtoets dan in het geval van retrieval practice via een oefentoets. Uit het bovengenoemde blijkt dat onderzoeksresultaten met betrekking tot het effect van retrieval practice op hogere orde kennis en vaardigheden niet eenduidig zijn, maar dat er ook positieve verwachtingen zijn.

1.2 Vraagstellingen en hypothesen

Op basis van bovengenoemde informatie is duidelijk geworden dat meer onderzoek nodig is naar de invloed van een oefentoets op de resultaten op een summatieve kennistoets met betrekking tot hogere orde kennis. Tevens kan dan gekeken worden of de soort feedback (enkel juist/onjuist feedback (categorie 1 feedback, zie Tabel 2)) en een oefentoets met toelichtende feedback (categorie 2 feedback, zie Tabel 2) hierbij een belangrijke rol kan spelen.

Dit resulteert in de centrale onderzoeksvraag: “Welk effect heeft het inzetten van een oefentoets, die qua inhoud en niveau vergelijkbaar is met de eindtoets, op de resultaten van de eindtoets met betrekking tot de hogere orde kennis? Meer specifiek adresseert dit onderzoek de volgende deelvragen:

Deelvraag 1 In hoeverre beïnvloedt het al dan niet afnemen van een oefentoets de scores op de eindtoets?

Deelvraag 2 In welke mate beïnvloedt de hoogte van de score op de oefentoets de score op de eindtoets?

Deelvraag 3 Wat is het effect van de feedbackvorm (juist/onjuist versus toelichtende feedback) bij de oefentoets op de eindtoetsscores?

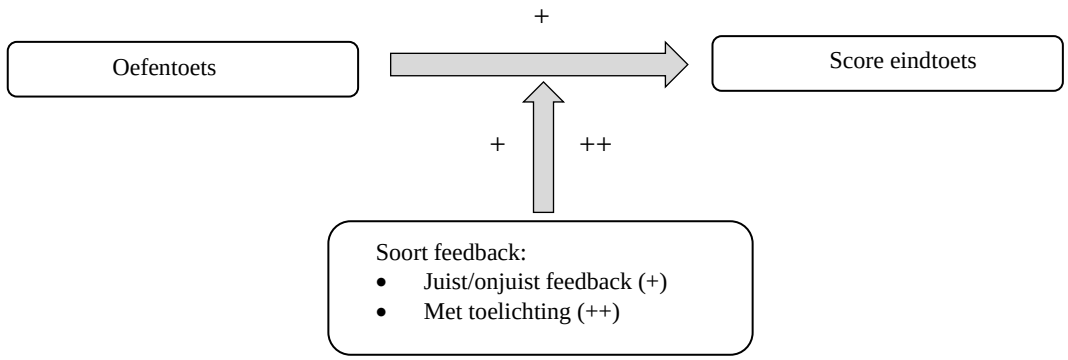
Deze deelvragen zijn op basis van het theoretisch kader vertaald naar de volgende hypothesen:

H1 Studenten die een oefentoets maken, scoren beter op de eindtoets met betrekking tot hogere orde kennis dan studenten die deze oefentoets niet maken.

H2 Studenten die een oefentoets maken met daarbij feedback in de vorm van een toelichting waarom het gegeven antwoord incorrect is (categorie 2), scoren beter op de eindtoets met betrekking tot hogere kennis dan studenten die een oefentoets maken met enkel feedback in de vorm van juist/onjuist (categorie 1).

Voor deelvraag 2 is geen hypothese opgesteld omdat de literatuur hieromtrent geen uitspraken doet.

In Figuur 2 staat het conceptueel model weergegeven waarin de onafhankelijke variabele (deelname aan de oefentoets), de modererende variabele (soort feedback bij de oefentoets) en afhankelijke variabele (de score op de eindtoets) zijn opgenomen.

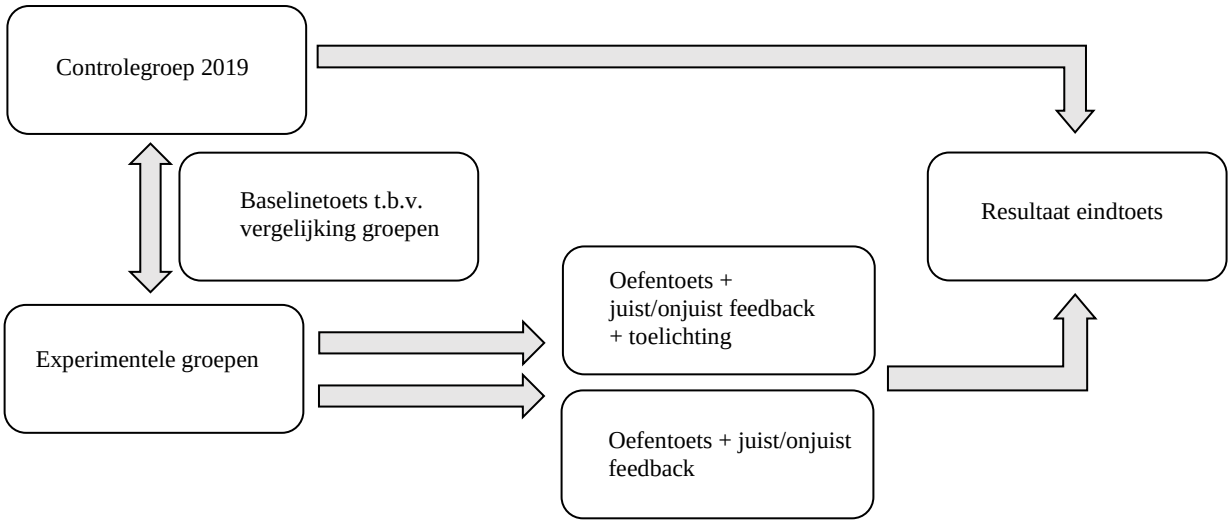


Figuur 2. Conceptueel model: de relatie tussen de oefentoets met verschillende vormen van feedback en de score op de eindtoets.

2. Methode

2.1 Ontwerp

Dit onderzoek richt zich op het effect van een oefentoets op de scores op een eindtoets. Om deze onderzoeksvraag te beantwoorden is een oefentoets ingezet met twee verschillende vormen van feedback. Om het causaal verband te kunnen vaststellen, is gekozen voor een experimenteel posttest only design met twee experimentele groepen en een controlegroep. Er is sprake van een zuiver experiment omdat de onderzoekspopulatie ad random toegewezen is aan een van de twee experimentele condities. Dit maakt het mogelijk verschillen te signaleren tussen de diverse condities waar deelnemers aan blootgesteld zijn. De invloed van onbekende variabelen wordt hiermee gereduceerd. Er is gekozen voor een posttest only design omdat een pretest waarin al een formatieve toets ingezet zou worden effect kan hebben op de resultaten op de eindtoets (testing-effect). Daarbij is het van weinig nut om gebruik te maken van een pretest omdat studenten nieuw zijn op de opleiding en over het algemeen de relevante voorkennis nog niet eigen hebben gemaakt. Om vast te stellen dat de experimentele groepen vergelijkbaar zijn met de controlegroepen is gebruik gemaakt van een baselinetoets. Het ontwerp is weergegeven in Figuur 3.

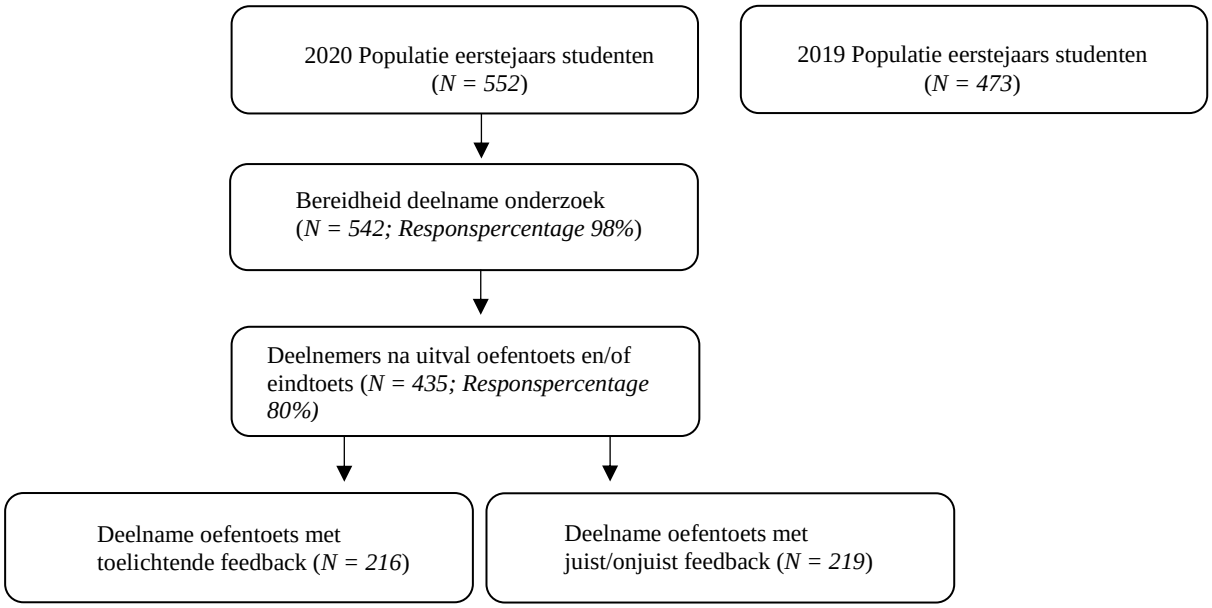


Figuur 3. Onderzoeksontwerp posttest only design met controlegroep.

2.2 Participanten

Voor dit onderzoek zijn propedeusestudenten uit studiejaar 2020 - 2021 van een hbo-opleiding Verpleegkunde benaderd als convenience sample. Het overgrote deel van deze studenten volgt de voltijdopleiding. Een klein deel de deeltijd- of duale opleiding. Alle eerstejaars studenten zijn aangeschreven om deel te nemen aan het onderzoek waarbij ze allen de oefentoets met betrekking tot de onderwijsinhouden ‘Kern’ en ‘Anatomie, Fysiologie, Pathologie (AFP)’ aangeboden kregen.

De reden voor het benaderen van alle studenten was de verwachting dat studenten voordeel zouden hebben van de interventie bij het maken van de eindtoets. Er is gekozen voor een controlegroep uit onderwijscohort 2019-2020. Deze controlegroep bestaat eveneens uit de gehele populatie eerstejaars studenten. Zij hebben inhoudelijk hetzelfde onderwijs gevolgd en hebben een vergelijkbare eindtoets gemaakt qua vorm en inhoud. De controlegroep heeft echter geen oefentoets gemaakt. Het zijn enkel hun resultaten op de baselinetoets en op de eindtoets die beschikbaar zijn gesteld door de opleiding. Deze gegevens zijn volledig anoniem aangeleverd. In 2019 hebben 473 studenten deelgenomen aan de summatieve eindtoets, in 2020 waren dit 510 studenten. Van de studenten uit cohort 2020 - 2021 die hebben aangegeven mee te willen doen met het onderzoek hebben 435 zowel de oefentoets als de eindtoets gemaakt. In Figuur 4 staan de gegevens van de controle- en experimentele groepen beschreven.



Figuur 4. Visuele weergave van aantal deelnemers in controlegroep en experimentele groepen

Alle docenten die lesgeven in beide modules zijn aangeschreven om mee te werken aan het onderzoek. Het betreft in totaal 25 docenten die aan 38 studiegroepen van gemiddeld 16 personen per groep lesgeven. Een a priori poweranalyse is uitgevoerd met behulp van *G*Power 3.1* om een voldoende steekproef te bepalen. Omdat de literatuur geen eenduidige uitspraak doet over de effectgrootte met betrekking tot hogere orde kennis, wordt uitgegaan van een *medium effect size* (Heinrich-Heine-Universität Düsseldorf, 2020). Met $f^2 = 0.5$, power van 95% en alpha van 0.05, was een totale steekproef van 76 nodig (Field, 2013; Heinrich-Heine-Universität Düsseldorf, 2019).

2.3 Materialen

2.3.1 Oefentoets (onafhankelijke en modererende variabele)

In samenwerking met docenten die ‘eigenaar’ zijn van Kern en AFP, is een twintigtal vragen opgesteld welke overeenkomen met de onderwerpen die besproken zijn in de lessen gedurende de eerste vier lesweken van Kern en AFP (zie Bijlage 1). De onderwerpen van Kern bestaan onder andere uit de volgende aspecten: CanMEDS-rollen in relatie tot diverse werkvelden, toepassen en analyseren van het verpleegkundig proces, positieve gezondheid gerelateerd aan verpleegkundig beroep en palliatieve zorg. In de module AFP staan de vaardigheden klinisch redeneren en het gebruik maken, toepassen en analyse van biomedische kennis centraal in de rol als zorgverlener. De werkgroep kennistoetsing heeft de vragen getoetst aan de hand van de toetsmatrijs en redigeerde de formulering van de toetsvragen zodanig dat de vormgeving, de inhoud en het niveau vergelijkbaar zijn met het type vragen dat in de eindtoets aan de orde komt. De oefentoets bestaat uit 20 vragen, tien vragen per onderdeel. De oefentoets A heeft als feedback juist/onjuist met daarbij toelichtende feedback (categorie 2 feedback) indien een vraag fout

beantwoord is en toets B heeft enkel als feedback juist/onjuist (categorie 1 feedback). Voorafgaand aan het maken van de toets hebben studenten van de eigen docent instructie ontvangen met betrekking tot het maken van de vragen en het verkrijgen van de feedback na elke vraag. Voor beide toetsen gold dat studenten na elke vraag actief de feedbackknop aan moesten klikken om de feedback te zien. Aan het einde van beide toetsen is, naar de student, teruggekoppeld hoeveel vragen in totaal correct zijn beantwoord inclusief het percentage goed beantwoorde vragen. Deze laatste feedback werd automatisch teruggekoppeld vanuit de software TestVision waarmee de toets digitaal wordt uitgezet. TestVision is een uitgebreid toetssysteem dat ondersteunend werkt bij het ontwikkelen, afnemen, nakijken, beoordelen en analyseren van toetsen. De betrokken hogeschool heeft een verwerkersovereenkomst met TestVision.

2.3.2 Posttest (afhankelijke variabele)

De posttest bestaat uit een kennistoets waarin de inhoud van zowel Kern als AFP worden getoetst. Voor beide experimentele groepen waren de toetsen identiek. De toets die de studenten uit de controlegroep hebben gemaakt, is van een vergelijkbaar niveau en inhoud. Om dit te borgen heeft de opleiding een heldere procedure voor het opstellen van de toets, met behulp van items uit de itembank. Aan de hand van een toetsmatrijs wordt de toets opgebouwd waardoor de representativiteit van de diverse onderwerpen evenwichtig verdeeld is en vergelijkbaar is over de jaren heen. De totale toets bestaat uit 80 juist/onjuist stellingen, 40 stellingen per module. De vragen op de eindtoets worden zoveel mogelijk op het niveau van toepassing en analyse gesteld. De studenten ontvingen na afloop een afgerond cijfer als de eindscore. De onderzoeker heeft daarnaast het percentage goed beantwoorde vragen ontvangen voor de gehele toets en voor de twee onderdelen afzonderlijk.

2.3.3. Baselinetoets

De baselinetoets is een opdracht die elk jaar in de eerste periode van de propedeuse gemaakt wordt door alle studenten. Het betreft een introductiemodule waarbinnen studenten in twee- of drietallen een vakblad schrijven voor een vooraf gekozen werkveld binnen de zorg (bijvoorbeeld algemene gezondheidszorg of geestelijke gezondheidszorg). Het doel is dat het begrip van belangrijke basisbeginselen van de verpleegkundige zorg aangetoond wordt, onder andere ethiek van zorg, positieve gezondheid en persoonsgerichte zorg. Aan het einde van de onderwijsperiode wordt dit vakblad ingeleverd. Voor zowel de controle- als experimentele groepen is deze toets met behulp van dezelfde rubric beoordeeld. Om de interbeoordelaarsbetrouwbaarheid te vergroten worden jaarlijks kalibreersessies georganiseerd. De groep die samengewerkt heeft aan het werkstuk ontvangt een gezamenlijk cijfer welke vervolgens per individuele student opgeslagen wordt in het studentvolgsysteem.

2.4 Procedure

Dit onderzoek werd goedgekeurd door de ethische commissie van de Open Universiteit (U/2020/01187/MQF) en de leidinggevende van de opleiding. Het onderzoek is uitgevoerd in lijn met de Algemene Verordening Gegevensbescherming. Bij de start van het studiejaar zijn de docenten geïnformeerd over het onderzoek en de rol die zij daarbij kunnen spelen. Zij hebben informatie ontvangen via de online docentenomgeving. Daarnaast hebben ze individueel een mail ontvangen met een toelichting over de werkwijze van dit onderzoek. Hierbij is gevraagd of ze bezwaar hebben tegen het afnemen van de oefentoets gedurende de les. Alle docenten, 25 in totaal, die de kernlessen verzorgen, hebben ingestemd met het afnemen van de oefentoets tijdens de kernles in onderwijsweek vijf. Voor het inzetten van de oefentoets is zoveel mogelijk aangesloten bij de ontwerprichtlijnen voor formatieve toetsen van Dirkx et al. (2019): 1. Het domein zou geschikt moeten zijn; 3. Constructive alignment is een uitgangspunt van de opleiding; 4. Er wordt alleen gebruik gemaakt van gesloten vragen; 5. Er is feedback beschikbaar; 6. De oefentoets wordt ingezet na een initiële leerfase; 7. Dezelfde stof wordt eenmaal formatief en eenmaal in de eindtoets getoetst; 8. De oefentoets en de eindtoets worden gespreid van elkaar aangeboden; 9. De oefentoets is niet vlak voor de eindtoets en 10. De oefentoets wordt actief door de docenten in het onderwijsprogramma ingezet. Dit onderzoek biedt mogelijk nieuwe inzichten ten aanzien van richtlijn 2 (formatieve toetsen moeten vooral gebruikt moeten worden voor onthouden, begrijpen, en toepassen van informatie), namelijk of de oefentoets ook nuttig is bij hogere orde kennis en vaardigheden. De oefentoets werd tijdens de les aangeboden in de veronderstelling dat dit de kans op deelname zou vergroten en om de condities zo vergelijkbaar mogelijk te houden. Nadat de afspraken met de docenten zijn gemaakt, hebben de studenten een nieuwsbericht ontvangen waarin ze zijn ingelicht over dit onderzoek. Op basis van studentnummer zijn studenten gerandomiseerd ingedeeld in experiment groep A of B. Beide varianten van de oefentoets konden hiermee binnen een studiegroep voorkomen. Participanten zijn niet toegewezen aan controlegroepen om ethische redenen: eerder onderzoek heeft namelijk aangetoond dat een formatieve toets positieve effecten kan hebben, hetgeen betekent dat studenten in een controlegroep nadeel zouden ondervinden ten opzichte van studenten die ingedeeld zijn in de experimentele groep. Als oplossing hiervoor zijn resultaten van studenten uit het vorige onderwijscohort (2019), toen geen oefentoets werd ingezet, als controlegroep gebruikt.

In studieweek vijf is de oefentoets aangeboden. In de inleidende tekst van deze toets stond het Informed Consent weergegeven. Indien studenten akkoord gaven voor deelname aan het onderzoek, vinkten ze dit aan op de introductiepagina van de toets. De enige persoonlijke informatie die hiervoor beschikbaar gesteld diende te worden was het studentnummer. Ook als studenten geen toestemming gaven, mochten zij de oefentoets maken. Voor de eindtoets gold de reguliere werkwijze zoals deze door de opleiding is ingeregeld. Studenten tekenen zich in voor de toets via de opleidingsportal. Zonder deze inschrijving kunnen studenten niet deelnemen aan de eindtoets. Deze toets is afgenomen in lesweek negen. Nadat de beoordelingen van deze toetsen zijn verwerkt door de ondersteunende dienst Toets en

Evaluatie Service Organisatie, zijn de resultaten teruggekoppeld aan de onderzoeker. Deze bestaan uit zowel de eindscore afgerond op een heel cijfer als het percentage goed beantwoorde vragen. De opleiding heeft een toetsanalyse uitgevoerd waardoor vragen verwijderd kunnen worden, wat gevolgen kan hebben voor de normering. Uitgaande van de percentagescores zorgt voor een betere vergelijking tussen de cohorten. Voorafgaand aan de analyse zijn de resultaten op de oefentoets en die van de posttest aan elkaar gekoppeld aan de hand van gepseudonimiseerde studentnummers. Na de koppeling zijn studentnummers verwijderd en was er sprake van een volledig geanonimiseerde dataset. De gegevens van de controlegroep zijn anoniem aangeleverd.

Indien studenten uit de experimentele groepen niet hebben deelgenomen aan de eind- of oefentoets, zijn deze uit het databestand verwijderd. Omdat deze niet meegenomen worden in het onderzoek is selectieve uitval een mogelijke bedreiging voor de representativiteit. Potentiele kenmerken van deze studenten in relatie tot de scores op de eindtoets, zijn niet in beeld gebracht omdat er geen verdere achtergrondgegevens zijn meegenomen.

In het experiment is aangenomen dat de populatie op de posttest vergelijkbaar is tussen de jaren 2019 en 2020. Ook is de inhoud van het vak identiek en het type toets dat wordt afgenomen vergelijkbaar. Om te controleren of deze aanname gerechtvaardigd is, zijn scores op een andere toets, vergelijkbaar voor de twee cohorten, met elkaar vergeleken. Dit betrof afgeronde rapportcijfers. Na afronding van het onderzoek zijn de belangrijkste bevindingen uit het onderzoek gerapporteerd via een nieuwsbericht aan de studenten en docenten.

2.5 Data-analyse

Het effect van een oefentoets op de resultaten op de eindtoets (deelvraag 1) is gemeten door de resultaten op de eindtoets van het onderwijscohort 2019 te vergelijken met de resultaten van de eindtoets bij de posttest in 2020. Ten behoeve van het kunnen beantwoorden van de tweede deelvraag is aan het onderwijscohort 2020, feedback op de oefentoets gegeven middels juist/onjuist of juist/onjuist met toelichting. De eerste stap was het maken van een univariate exploratieve analyse per groep en voor de groepen samen, waarbij gekeken is naar het gemiddelde en de spreiding van de posttestscore. Een van de redenen om deze analyse uit te voeren, was het vinden van uitschieters.

Ten behoeve van de assumpties van ANOVA is de normaalverdeling van de steekproefverdeling gecontroleerd met behulp van de Kolmogorov-Smirnov Test (geschikt voor grote steekproeven > 50). Aanvullend zijn Q-Q-plots opgevraagd. Volgens de Centrale Limiet Theorie geldt dat aangenomen kan worden dat aan normaliteit is voldaan op het moment dat de steekproef een grote omvang heeft, ook al is de variabele in de populatie zelf niet normaal verdeeld (Field, 2013). De Levene's test is gebruikt om de homogeniteit van de onafhankelijke variabele te bepalen. Het aangehouden significantieniveau is $p \leq .05$. Dit betekent dat de nulhypothese wordt verworpen als de alpha kleiner is dan 5% of als 0 buiten het 95%-betrouwbaarheidsinterval voor het groepseffect valt. In het geval van een significant verschil

tussen de groepen, zijn vervolgens *planned contrasts* uitgevoerd omdat er sprake is van gerichte hypotheses. Deze werkwijze maakt de kans op type I-fouten kleiner (Field, 2013).

Ter controle van de assumpties van regressie zijn P-P plots gemaakt voor de normaalverdeling van de residuen. Voor de controle van de homoscedasticiteit zijn scatterplots gemaakt. Alle analyses zijn uitgevoerd met SPSS (versie 26).

In het experiment is aangenomen dat de populatie op de posttest vergelijkbaar is tussen studiejaren 2019 en 2020. Ook is de inhoud van het vak identiek en het type toets dat wordt afgenomen onderling vergelijkbaar. Om deze aanname te verifiëren, zijn scores op een baselinetoets meegenomen. Zowel in 2019 als in 2020 is deze toets gemaakt. De vergelijking is geanalyseerd met behulp van ANOVA en aanvullend met de non-parametrische Kruskal-Wallis test.

Voor de eerste onderzoeksvraag "In hoeverre beïnvloedt het al dan niet afnemen van een oefentoets de scores op de eindtoets?" en de derde onderzoeksvraag "Wat is het effect van de feedbackvorm (juist/onjuist versus toelichtende feedback) bij de oefentoets op de eindtoetsscores?" is gekeken naar het verschil in gemiddelde posttestresultaten (afhankelijke variabele) tussen de drie onderzoeksgroepen waarbij de scores op de baselinetoets als covariaat zijn meegenomen. Als er individuele waarnemingen zijn die extreem scoren, is het van belang te begrijpen wat er aan de hand is en de achtergrond ervan proberen te begrijpen om vervolgens te beslissen of ze in het databestand blijven of worden verwijderd. Gezien het feit dat er gebruik is gemaakt van percentagescores en er geen open antwoordvelden waren, werden geen uitschieters verwacht. Van de scores op de posttest is, naast de beschrijvende samenvatting, ook een histogram gemaakt om een idee te krijgen van de verdeling van de verschillende variabelen. Per groep is een boxplot gedraaid van de scores om de groepen te vergelijken. Op basis van ANCOVA, met de scores op de baselinetoets als covariaat, en bijbehorende post-hoc testen is bepaald of er significante verschillen zijn tussen de onderzoeksgroepen. Om te achterhalen of er een relatie bestaat tussen hoogte van de score op de tussentoets met de resultaten op de eindtoets, is de tussentoets als covariaat meegenomen (onderzoeksvraag 2). Vervolgens is gekeken of de score op de oefentoets van invloed is op de eindtoetsscore met behulp van de Pearson's correlatie. Een correlatiecoëfficiënt van + 1 laat een perfect positief verband zien, een coëfficiënt van - 1 een perfect negatief verband en een coëfficiënt van 0 toont aan dat er geen enkel verband is tussen de twee variabelen. Als er sprake was van een positieve of negatieve correlatie, is aanvullend een enkelvoudige regressieanalyse uitgevoerd om na te gaan in hoeverre de oefentoets de variantie in de scores op de eindtoets verklaart.

3. Resultaten

De centrale vraag die beantwoord wordt in dit onderzoek is "Welk effect heeft het inzetten van een oefentoets, die qua inhoud en niveau vergelijkbaar is met de eindtoets, op de resultaten van de eindtoets met betrekking tot de hogere orde kennis?". Om te meten of de interventie, in de vorm van een

oefentoets, effect heeft, is een aantal analyses uitgevoerd. In alle gevallen is het aangehouden significantieniveau $p \leq .05$. De controlegroep bestond uit 473 studenten en de totale experimentele groep uit 435. Deze laatste groep is verdeeld over 216 studenten die de oefentoets gemaakt hebben met toelichtende feedback en 219 studenten die de oefentoets gemaakt hebben met enkel juist/onjuist feedback. Voor de assumptie dat de controlegroep vergelijkbaar is met de experimentele groep zijn resultaten van een baselinetoets gebruikt. In Tabel 3 staan de kengetallen van deze baselinetoets weergegeven.

Tabel 3

Kengetallen ten behoeve van vergelijkbaarheid van onderzoeksgroepen

	Score baselinetoets		
	n	*M*	*SD*
Experimentele groep A (toelichtende feedback)	182	7.80	1.03
Experimentele groep B (juist/onjuist feedback)	194	7.77	.98
Controlegroep	345	7,45	.88

De scores op de baselinetoets voor de controlegroep zijn niet normaal verdeeld, $D(345) = .242$, $p < .001$. Dit geldt ook voor experimentele groep A (toelichting), $D(182) = .274$, $p < .001$ en voor experimentele groep B (juist/onjuist), $D(194) = .267$, $p < .001$. De Q-Q plots en de histogrammen laten echter zien dat er sprake is van een (linksscheve) normaalverdeling. Aangezien het hier schoolcijfers betreft, is dat wel een verwachte scheefheid. Scores onder de vijf worden veel minder behaald. Er is een significant verschil in de gemiddelde scores op de baselinetoets van de drie groepen $F(2.718) = 11.658$; $p < .001$. De post-hoc-Tukey-toets toont significante verschillen tussen experimentele groep A (toelichtende feedback) en de controlegroep ($p < .001$) en experimentele groep B (juist/onjuist) en de controlegroep ($p = < .001$). Er is geen significant verschil in de scores op de baselinetoets tussen beide experimentele groepen ($p = .953$). Omdat niet geheel voldaan is aan de assumptie van de normale verdeeldheid, is als aanvulling non-parametrisch getoetst met behulp van de Kruskal-Wallis test. Deze liet eenzelfde beeld zien ($H(2) = 27.36$, $p < .001$). De follow-up scores bevestigden de resultaten uit de ANOVA. Dit betekent dat de controlegroep niet als gelijk beschouwd kan worden aan de experimentele groepen en dat de vergelijking van de eindscores gecorrigeerd moeten worden voor dit verschil.

Tabel 4 laat respectievelijk het aantal deelnemers, gemiddelde scores op de oefentoets en eindtoets en spreiding van deze scores voor de verschillende onderzoeksgroepen zien (*n, M, SD*). Omdat

er sprake is van een beperkte uitval is de verwachting dat de representativiteit van de resultaten geborgd is gebleven.

Tabel 4

Deelnemers, gemiddelden en standaarddeviaties voor de interventie en nameting bij de experimentele en controlegroepen

	Interventie			Nameting		
	n	*M*	*SD*	*n*	*M*	*SD*
Experimentele groep A (toelichtende feedback)	216	71.85	14.25	216	74.82	7.68
Experimentele groep B (juist/onjuist)	219	77.35	15.39	219	73.61	7.31
Controlegroep 2019	-			473	74.23	7.54

De eerste deelvraag, “In hoeverre beïnvloedt het al dan niet afnemen van een oefentoets de scores op de eindtoets?” kan beantwoord worden met ANCOVA. Omdat de controlegroep, op basis van de scores op de baselinetoets, niet als gelijk beschouwd kan worden aan de experimentele groepen in de analyse zijn de scores op de baselinetoets als covariaat meegenomen. Hiervoor is gecontroleerd of aan de assumpties van de ANCOVA is voldaan. De varianties voor de verschillende onderzoeksgroepen zijn gelijk. De Levene’s test is niet significant (p = >.05) en ook aan de assumptie van gelijke regressiecoëfficiënten is voldaan (p = .685). Experimentele groep A (toelichtende feedback) scoorde hoger op de eindtoets (M = 74.82) dan experimentele groep B (juist/onjuist) (M = 73.61) en de controlegroep (M = 74.23). Uit de ANCOVA bleek dat de covariaat significant de scores op de eindtoets voorspelt (F = 1.715) = 32.455, p = <.001. Echter, zelfs wanneer de covariaat meegenomen wordt, blijkt dat de scores op de eindtoets niet significant verschillend zijn voor de drie groepen F(1.715) = .406, p = .667. In dit onderzoek is daarmee geen effect aangetoond van het inzetten van een oefentoets op de resultaten op de eindtoets.

Voor de tweede deelvraag, “In welke mate beïnvloedt de hoogte van de score op de oefentoets de score op de eindtoets?” is bekeken met behulp van ANCOVA, waarbij de scores op de oefentoets als covariaat zijn meegenomen. De reden hiervoor is een significant verschil in resultaten op de oefentoets tussen beide experimentele groepen (F (1.433) = 14.949, p <.001). Uit de ANCOVA bleek dat de interactie tussen de onafhankelijke variabele (experimentele groepen) en de covariaat (scores oefentoets) significant van elkaar verschillen F(1,431) = 8.174, p = .004. Dit houdt in dat de score op een oefentoets voor beide groepen een ander effect heeft op de resultaten op de eindtoets. Er wordt hiermee niet voldaan

aan de assumptie van gelijke regressie coëfficiënten waardoor de scores op de oefentoets niet als covariaat kunnen worden meegenomen in de analyses.

Om te beoordelen of de hoogte van de score op de oefentoets de resultaten op de eindtoets beïnvloedt, is de Pearson's correlatie berekend. Voor beide experimentele groepen samen geldt dat er sprake is van een significante, lage positieve correlatie tussen de scores op de oefentoets en de eindtoetsscore ($r = .19$; $p < .001$). Om te bepalen of deze significante correlatie voor beide experimentele groepen geldt, is tevens een correlatie berekend. Voor de groep die de oefentoets maakte met toelichtende feedback bestaat een significante, lage positieve correlatie tussen de scores op de oefentoets en de eindtoetsscore ($r = .33$; $p < .001$). Voor de groep die de oefentoets heeft gemaakt met juist/onjuist feedback wees uit dat er een niet-significant, zwak verband bestaat tussen oefentoets en eindscore $r = .08$; $p = .225$. Voor de experimentele groep waarbij de correlatie significant bleek, is een enkelvoudige regressieanalyse gedaan. De scores op de oefentoets blijkt een significante voorspeller van de scores op de eindtoets. De score op de oefentoets verklaart een significant deel van de variantie in eindtoetsscores ($R^2 = .109$; $F(1.214) = 26.204$; $p < .001$). Dit betekent dat van de variantie in eindtoetsscore 10.9% verklaard wordt door de score op de oefentoets.

De derde onderzoeksvraag, "Wat is het effect van de feedbackvorm (juist/onjuist versus toelichtende feedback) bij de oefentoets op de eindtoetsscores?" is getoetst met behulp van ANOVA. Ondanks dat de resultaten op de oefentoets bij beide experimentele groepen significant van elkaar verschillen (F (1.433) = 14.949, p <.001), zijn de resultaten op de eindtoets tussen de groep die toelichting kreeg en de groep die feedback kreeg in de vorm van juist/onjuist niet significant verschillend van elkaar (F (1.433) = 1.874, p = .172). Het onderzoek toont hiermee niet aan dat het type feedback van invloed is op de eindtoetsscores.

4. Discussie en conclusie

Dit onderzoek richtte zich op de vraag in hoeverre de resultaten op de eindtoets beïnvloed worden door het maken van een oefentoets die qua inhoud en niveau vergelijkbaar is met de eindtoets. Tijdens het experiment kregen studenten een oefentoets aangeboden, met juist/onjuist feedback of toelichtende feedback. De controlegroep kreeg deze interventie niet aangeboden. De belangrijkste bevindingen op basis van de gevonden resultaten zijn dat (1) het aanbieden van een oefentoets niet leidde tot significant hogere scores op de eindtoets, (2) de score op de oefentoets een significante, maar lage positieve correlatie laat zien met de score op de eindtoets, dit geldt echter alleen voor de experimentele groep die toelichtende feedback ontving, voor de andere groep was deze correlatie niet significant, en (3) er geen significant verschil zichtbaar was tussen de verschillende vormen van feedback op de resultaten op de eindtoets. In onderstaande paragraaf worden deze bevindingen achtereenvolgens besproken in relatie tot bestaande literatuur.

4.1 Effect van een oefentoets op resultaten eindtoets

De eerste hypothese stelde dat de scores op de eindtoets hoger zouden zijn als studenten een oefentoets maakten. De univariate analyse (ANOVA) laat zien dat deze stelling niet bevestigd wordt. Er is geen sprake van een significant verschil tussen de controlegroepen die geen oefentoets gemaakt hebben en de experimentele groepen. De experimentele groep die de oefentoets maakte met toelichtende feedback scoorde absoluut gezien wel het hoogste, echter niet significant beter dan de andere groepen. De andere experimentele groep scoorde het laagst, maar ook dat was niet significant. Bovenstaande bevindingen weerspreken de bestudeerde literatuur en eerdere onderzoeken die hebben plaatsgevonden. Onder andere Roediger en Karpicke (2006) laten zien dat het testing effect ervoor kan zorgen dat de resultaten op de eindtoets hoger zijn. Het toetsen tussen leermomenten heeft vooral een effect op het onthouden van de lesstof op de lange termijn. Dat effect is het sterkst als lerenden niet alleen opnieuw doornemen wat zij niet wisten, maar vooral als zij alles opnieuw bestuderen (d & Sweller, 2015). Vanuit het voorliggende onderzoek kan niet achterhaald worden in hoeverre studenten daadwerkelijk de gehele leerstof opnieuw hebben bestudeerd of dat ze zich hebben beperkt tot de nieuwe leerstof die na de tussentoets is behandeld. Dat is een onzekere factor in het onderzoek.

Een mogelijke verklaring dat het effect niet zichtbaar was in het voorliggende onderzoek is het feit dat niet de gehele lesstof is bevraagd. Butler, Karpicke, en Roediger (2007) en Yeo en Fazio (2019) laten namelijk zien dat een mogelijk risico van het testing effect is dat de inhoud van de tussentoets, dat als voorbeeld kan dienen voor de student, beter wordt onthouden dan leerstof dat niet in de tussentoets is opgenomen. Ondanks dat er zoveel mogelijk is aangesloten bij de wetenschappelijke richtlijnen voor het vormgeven van formatieve toetsing, is het mogelijk dat niet alle richtlijnen even adequaat aan bod zijn gekomen. Richtlijn vier geeft aan dat het van belang is dat er bij een oefentoets gebruik gemaakt moet worden van zowel open als gesloten vragen. In dit onderzoek is er enkel gebruik gemaakt van gesloten vragen. Bij het beantwoorden van, in dit geval, een vraag waarop het antwoord juist of onjuist is, wordt voornamelijk een beroep gedaan op herkenning dan op actievere retrieval practice die bij open vragen vereist wordt. Als er feedback wordt toegevoegd, wordt een deel van de beperkte actieve retrieval gecompenseerd. Echter, uit onderzoek blijkt dat een tussentoets waarbij de combinatie gemaakt wordt van open en gesloten vragen het meeste effect laten zien (Adesope et al., 2017; Pan & Agarwal, 2018). Een andere richtlijn betreft de 20% vuistregel (Dirkx et al., 2019). Dit houdt in dat de inzet van een eerste oefentoets na 20% van de behandelde stof het meest efficiënt is. Dit betreft een regel die opgaat in het geval van kleine tijdintervallen, bijvoorbeeld een eindtoets tien dagen na de start van een onderwijseenheid. In de module die als uitgangspunt is genomen in dit onderzoek geldt een tijdsinterval van negen weken. De vraag is dus of één enkele toets voldoende is gebleken om de scores op de eindtoets te beïnvloeden. De reviewstudies van Adesope et al. (2017) en Rowland (2014) tonen aan dat een keer tussentijds toetsen een even groot of zelfs een groter effect heeft op de eindtoets dan dat er meerdere keren tussentijds getoetst wordt. Karpicke en Roediger (2008), Pyc en Rawson (2011) en Roediger en

Karpicke (2006) laten echter zien dat meerdere toetsmomenten de vergeetcurve minder snel laat dalen. Tegelijkertijd neemt de toegevoegde waarde van elke toets wel af. De combinatie van slechts een enkele toets en het grote tijdsinterval tussen de oefentoets en de eindtoets in dit onderzoek kan mogelijk bijgedragen hebben aan het feit dat er geen significante verschillen zichtbaar zijn tussen de controle- en experimentele groepen.

4.2 Effect van score oefentoets op resultaat eindtoets

De tweede onderzoeksvraag was "In welke mate beïnvloedt de hoogte van de score op de tussentoets de score op de eindtoets?" is gesteld omdat er weinig tot geen literatuur gevonden is met betrekking tot dit onderwerp. Dit wekte bij de onderzoeker interesse op. Opvallend was dat er wel veel gesproken werd over het effect van het afnemen van een oefentoets (pure-effecten) maar dat er niet of nauwelijks is onderzocht in hoeverre de resultaten op een oefentoets leiden tot betere scores op de eindtoets. Daarom is voor deze onderzoeksvraag geen hypothese opgesteld. Uit de resultaten blijkt dat er een lage maar significante relatie lijkt te bestaan tussen de score op de oefentoets en de resultaten op de eindtoets bij studenten uit de experimentele groep die toelichtende feedback hebben gekregen. Opvallend hierbij is dat het effect van de tussentoets op de resultaten op de eindtoets significant is voor de groep die lager scoorde op de tussentoets, maar wel (niet) significant hoger scoorde op de eindtoets. Dit zou kunnen betekenen dat de hoogte van de scores op de tussentoets niet per se een bijdrage leveren aan de resultaten op de eindtoets, maar dat het om het testing effect gaat zoals uit onderstaande bronnen blijkt.

In veel onderzoek, onder andere dat van Roediger en Karpicke (2006) ligt de nadruk op het testing effect waarbij het voornamelijk gaat om het proces van het actief ophalen van de informatie uit het geheugen. Het testing effect heeft met name betrekking op de 'pure effecten' van toetsing waar het voorliggende onderzoek de nadruk op legt. Ook als er wat breder gekeken wordt, bijvoorbeeld naar resultaten uit onderzoek naar post-effecten van toetsing, ligt de focus op het effect van het maken van de toetsen en niet zozeer op de mate van het succesvol maken ervan. Een illustratief onderzoek is het onderzoek van Gijbels, van de Watering, en Dochy (2005). Zij stelden zichzelf de vraag of het doorlopen van oefentoetsen een positief effect had op de resultaten op de eindtoets. Het betrof een onderzoek waarbij studenten meerdere oefeningen maakten waarop tussentijds feedback gegeven werd door de docent. De resultaten lieten zien dat studenten die alle tussentijdse oefeningen gemaakt hadden, aanzienlijk beter scoorden op de eindtoets dan studenten die deze toetsen niet of niet allemaal gemaakt hadden. Ook in dit onderzoek is niet expliciet gekeken naar de samenhang tussen de het succesvol maken van de tussentijdse oefeningen en de resultaten op de eindtoets.

Ten slotte is het interessant te kijken naar mogelijke verklaringen voor de verschillen in scores op de tussentoets tussen beide onderzoeksgroepen in het voorliggende onderzoek. Omdat de studenten ad random zijn toegewezen aan een van de onderzoeksgroepen en de groepen een behoorlijke omvang hebben, is niet te verwachten dat de omstandigheden waaronder de toets gemaakt is of dat de

samenstelling van de groepen dermate verschillend zijn dat hiermee de verschillen verklaard kunnen worden. Ook de scores op de baselinetoets voor beide groepen zijn niet significant verschillend van elkaar. Dit betekent dat het, op basis van de gekozen aanpak en de verzamelde gegevens niet inzichtelijk is of de gerapporteerde verschillen aan toeval toegeschreven kunnen worden.

4.3 Invloed type feedback op resultaat eindtoets

De tweede hypothese stelde dat het type feedback van invloed zou zijn op de resultaten op de eindtoets. Er werd verondersteld dat feedback met toelichting meer effect zou hebben dan enkel juist/onjuist feedback. De resultaten van ANOVA toont deze aanname niet aan. Er is geen significant verschil tussen beide experimentele groepen. De eindscore voor de groep met toelichtende feedback op de items van de oefentoets is lager maar dit verschil is niet significant. De scores op de tussentoets verschillen wel significant van elkaar, waarbij de experimentele groep met beperkte feedback hoger scoort. Dit is niet terug te zien in de eindtoets. Deze groep scoort absoluut gezien het laagst van alle groepen, zij het niet significant. De verschillen zijn dermate klein, waardoor een verklaring hiervoor niet voorhanden is. Het kan puur toeval zijn. Er is in dit onderzoek gekozen voor feedback na elke vraag. Roediger en Butler (2011) geven aan dat het bij gesloten vragen erg belangrijk is om feedback te geven omdat er anders misconcepties kunnen ontstaan in verband met herkenning van de verkeerde antwoorden. Volgens Brookhart dienen misconcepties bij lagere orde kennis en vaardigheden zo snel mogelijk gecorrigeerd te worden (2009). De timing van de feedback blijkt echter ook van belang te zijn. Volgens Butler et al. (2007) is voornamelijk de timing van feedback van belang voor de prestaties op de eindtoets. Feedback dat direct na elke vraag wordt getoond levert minder effect op dan dat deze feedback op een later moment gedeeld wordt. Uit onderzoek blijkt echter dat uitgestelde feedback, aan het einde van een toets, meer effect teweegbrengen (Roediger & Butler, 2011). In het onderhavige onderzoek is het verschil tussen categorie 1 en 2 feedback onderzocht en is categorie 3 met het feedforward element niet meegenomen. Wellicht had het inzetten van deze vorm van feedback tot andere inzichten kunnen leiden.

4.4 Hogere orde kennis

In dit onderzoek is onderzocht in hoeverre de oefentoets van nut is voor hogere orde kennis. Uit de resultaten blijkt dat het effect van de oefentoets niet significant is op de eindtoetsresultaten.

De vraag is echter of de toets voldoende onderscheid maakt tussen lagere en hogere orde kennis en in hoeverre de vormgeving van de huidige oefentoets en eindtoets, voldoende mogelijkheden biedt om hogere orde kennis te toetsen (juist/onjuist stellingen). De herziene taxonomie van Bloom (Anderson & Krathwohl, 2001) maakt een onderscheid tussen vier typen kennis zien, te weten; feitelijke, conceptuele, procedurele en metacognitieve kennis die vervolgens op de niveaus onthouden, begrijpen, toepassen, evalueren en creëren bevraagd kan worden. In samenspraak met docenten is, ten behoeve van dit onderzoek, gekozen voor een kennistoets waarin aangenomen wordt dat het hogere orde kennis bevraagt. De toets heeft met name betrekking op feitelijke en conceptuele kennis. Van de studenten in

het onderzoek wordt verwacht dat ze basisbegrippen en feiten beheersen die nodig zijn om het vak van verpleegkundige te beheersen. Daarnaast is het van belang dat ze aantonen inzicht te hebben in theorieën en modellen. De toets bestond uit 80 juist/onjuist stellingen. Ondanks dat er opvattingen bestaan dat dit type vraag minder geschikt is voor het bevragen van hogere orde kennis, zijn Ebel en Frisbie (1991) van mening dat alle kennis, dus ook van hogere orde, uit te drukken is in proposities, bijvoorbeeld een stelling of uitspraak die juist of onjuist is. Het blijkt namelijk dat de vraagvorm minder belangrijk is dan de vraaginhoud. De vraag is, door de hoge gokkans, of er een voldoende duidelijk beeld ontstaat bij de beoordelaar over de beheersing van de leerstof (Van Berkel et al., 2017). De invloed van ad random gokken neemt af als het aantal items in een toets toeneemt. Dit is in de toets uit dit onderzoek ondervangen door voldoende vragen op te nemen in de toets en de cesuur boven de random gokkans te leggen. Een groot deel van de kennis dat de student moet aantonen bevindt zich op het niveau van toepassen en in een enkel geval op het niveau van begrijpen (lagere orde) en evalueren (hogere orde). Zowel Anderson en Krathwohl (2001) als Brookhart (2010) scharen onthouden en begrijpen onder de lagere orde kennis en evalueren en creëren onder de hogere orde kennis. Over het toepassen van kennis zijn de meningen verdeeld. Waar Brookhart dit als hogere orde beschouwt, zien Anderson en Krathwohl (2001) dit als een lagere orde vaardigheid. Het toepassen van kennis houdt in dat de lerende feiten en concepten gebruikt om nieuwe of vergelijkbare problemen uit het verleden, op te lossen. Het betreft veelal problemen die een eenduidig antwoord verlangen (Brookhart, 2010). Door deze verschillende opvattingen vanuit de wetenschap is er discussie mogelijk over de focus van het voorliggende onderzoek. Volgens de opvatting van Brookhart (2010) heeft dit onderzoek daadwerkelijk betrekking op hogere orde kennis waar Anderson en Krathwohl (2001) het zouden beschouwen als een onderzoek naar lagere orde kennis. In beide gevallen is de conclusie hetzelfde, namelijk dat er geen effect van een oefentoets is aangetoond.

Concluderend kan gesteld worden dat dit onderzoek geen significant effect laat zien van het inzetten van een oefentoets op de resultaten van een eindtoets met betrekking tot hogere orde kennis. Er bleek ook geen verschil te zijn tussen de vorm waarin de feedback aangeboden werd. De aanname dat dit wel van belang is wordt in dit onderzoek ter discussie gesteld.

4.5 Beperkingen en aanknopingspunten voor vervolgonderzoek

De voornaamste sterkte van de onderhavige studie is het feit dat het een aanvulling biedt op de beperkt beschikbare bestaande literatuur naar de invloed van oefentoetsen op de beheersing van hogere orde kennis- en vaardigheden op een eindtoets. Agarwal (2018) en de overzichtsstudie van Adesope et al. (2017) laten zien dat er bewijzen lijken te zijn dat oefentoetsen ook van nut zijn voor hogere orde kennis, maar dat er aanvullend onderzoek nodig is.

Naast de waarde van dit onderzoek, brengt het ook beperkingen met zich mee. Door de opzet van het onderzoek is de invloed van zoveel mogelijk andere onafhankelijke variabelen uitgesloten. Echter, het onderzoek heeft in een realistische onderwijssetting plaatsgevonden, waardoor het alsnog mogelijk

is dat andere variabelen, buiten de interventie om, een rol hebben gespeeld bij de resultaten op de eindtoets (Creswell, 2014). De controlegroep bleek niet geheel vergelijkbaar te zijn met de experimentele groepen, gebaseerd op een baselinetoets. De vraag is of het bekijken van een enkele vergelijkbare toets, voldoende aanknopingspunten biedt om een uitspraak te doen over het al dan niet vergelijkbaar zijn van de controle- en experimentele groepen. Dit neemt niet weg dat het gebruik maken van een controlegroep uit een ander cohort risico's met zich meebrengt. Een belangrijk onderscheid tussen de twee groepen is bijvoorbeeld de vormgeving van het onderwijs. De studenten uit de experimentele groepen volgden, in verband met COVID-19, in tegenstelling tot de controlegroep volledig onlineonderwijs. Ondanks dat de lesinhouden en de docenten vergelijkbaar zijn, is het reëel dat dit invloed heeft op de voorbereiding van de studenten op de eindtoets. Dit kan een positieve impact hebben gehad, namelijk minder afleiding tijdens de gedeeltelijke lockdown waardoor er meer tijd besteed is aan studeren. Het kan echter ook de studie negatief beïnvloed hebben, bijvoorbeeld door een lagere betrokkenheid bij de opleiding door minder contact met medestudenten en docenten.

Een andere beperking betreft de oefentoets. Studenten dienden na elke vraag actief de link naar de feedback aan te klikken. Hierin stond of de vraag goed of fout beantwoord was én, in het geval van onderzoeksgroep A, de informerende feedback. Docenten hebben de studenten hierop gewezen maar of ze dat ook daadwerkelijk aangeklikt hebben, was na afloop niet verifieerbaar. Dit betekent dat de resultaten bij de onderzoeksvraag over de verschillende vormen van feedback moeilijk te verifiëren zijn.

Voor vervolgonderzoek is aan te bevelen om het onderzoek zodanig te organiseren dat zowel de experimentele als controlegroepen afkomstig zijn vanuit één onderwijscohort. Hiermee is de kans kleiner dat er belangrijke verschillen tussen de groepen de resultaten van het onderzoek beïnvloeden. Daarbij is het aan te bevelen om studenten tijdens het onderzoek naar hun studiegedrag te bevragen om inzicht te krijgen in mogelijke factoren die de scores op de eindtoets beïnvloeden.

Ondanks de behoorlijke omvang van de onderzoekspopulatie is het onderzoek binnen een enkele opleiding voor één type toets uitgevoerd. Het is interessant te kijken of een vergelijkbaar onderzoek bij andere type toetsen en andere studentenpopulaties van bijvoorbeeld andere studierichtingen of studiejaren, vergelijkbare resultaten opleveren.

4.6 Reflectie op wijzigingen van het onderzoek als gevolg van coronamaatregelen

Een belangrijke kanttekening is de wijziging van de opzet van het onderzoek door COVID-19. In maart 2020 stond een ander onderzoek in de planning. De vraagstelling was grotendeels hetzelfde maar de uitvoering was zowel bij een andere doelgroep (toekomstige docenten in het voortgezet onderwijs) als voor een andere type toets (het schrijven van een zelfreflectie na afloop van een stageperiode). Op het moment dat het onderzoek uitgezet zou gaan worden, kwam de lockdown en zag zowel de stage als het onderwijs er anders uit dan ten tijde van het maken van het onderzoeksvoorstel. De vele wijzigingen die dit met zich meebracht, heeft de betreffende opleiding en de onderzoeker ertoe doen besluiten het onderzoek niet door te laten gaan. Door de onzekerheid die bleef bestaan is voor een pragmatische

oplossing gekozen, waardoor het meten van de hogere orde kennis op een andere wijze heeft plaatsgevonden dan gepland. In het oorspronkelijke onderzoek stond een reflectie-opdracht als toets gepland en in de aangepaste opdracht is dit een digitale kennistoets geworden. Ondanks dat deze toets bedoeld is als hogere orde kennistoets, is de veronderstelling dat de hogere kennis beter met de eerder geplande reflectietoets gemeten had kunnen worden. Zie daarvoor ook de opmerkingen in paragraaf 4.4. Daarnaast moest het onderzoek binnen een beperkt tijdsbestek kunnen plaatsvinden om een zo klein mogelijk risico te lopen om opnieuw het onderzoek te moeten afbreken. Dit heeft ervoor gezorgd dat het niet mogelijk was om binnen het onderwijscohort dat deelnam aan het onderzoek een controlegroep samen te stellen, maar dit is opgelost door een controlegroep uit eerdere cohorten samen te stellen.

Referenties

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659-701. doi:10.3102/0034654316689306

Agarwal, P. K. (2018). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology, 111*(2), 189-209. doi:10.1037/edu0000282

Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.

Assessment Reform Group (2002). *Assessment for learning: 10 principles. Research-based principles to guide classroom practice.* Opgehaald op 12 oktober 2019, van http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/doc/Afl_principles.pdf

Barahal, S. L. (2008). Thinking about thinking. *Phi Delta Kappan, 90*(4), 298-302.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principle, policy & Practice, 5*(1), 7-74. doi:10.1080/0969595980050102

Brookhart, S. M. (2007). Expanding views about formative assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into Practice* (pp. 43-62). New York: Teachers College Press.

Brookhart, S. M. (2010). *How to Access Higher-order Thinking Skills in Your Classroom*. Alexandria: ASCD.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118-1133. doi:10.1037/a0019902

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*(4), 273-281. doi: 10.1037/1076-898x.13.4.273

Creswell, J. W. (2014). *Educational research: Planning, conducting and evaluating quantitative and qualitative research.* Essex: Pearson Education Limited.

Dirkx, K., Joosten-Ten Brinke, D., & Camp, G. (2019). *Ontwerprichtlijnen voor formatief toetsen vanuit de geheugenpsychologie 1 + 1 = 3?*. Geraadpleegd op 14 oktober 2019, van https://www.nro.nl/wp-content/uploads/2019/05/Eindrapport-405-17-711_Definitief.pdf

Dobson, J., Linderholm, T., & Perez, J. (2018). Retrieval practice enhances the ability to evaluate complex physiology information. *Medical Education, 52*(5), 513–525. doi:10.1111/medu.13503

Dochy, F., Segers, M. S. R., Gijbels, D., & Struyven, K. (2007). Assessment engineering: Breaking down barriers between teaching and learning, and assessment. In D. Boud, & N. Falchikov

(Eds.), *Rethinking assessment in higher education* (pp. 87-101). Routledge/Taylor & Francis Group.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques. *Psychological Science in the Public Interest, 14*(1), 4-58. doi:10.1177/1529100612453266

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement.* Prentice Hall.

Heinrich-Heine-Universität Düsseldorf. (2019). *Allgemeine Psychologie und Arbeitspsychologie G*Power: Statistical Power Analyses for Windows and Mac*. Geraadpleegd op 5 november 2020, van http://www.gpower.hhu.de/

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical poweranalysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. Geraadpleegd op 4 juli 2020, van https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). London: SAGE Publications Ltd.

Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigation frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology, 109*(8), 1067-1083. doi:10/1037/edu0000197

Gijbels, D., van de Wetering, G., & Dochy, F. (2005). Integrating assessment tasks in a problem-based learning environment. *Assessment and Evaluation in Higher Education, 30*(1), 73-86. doi: 10.1080/0260293042003243913

Glover, C., & Brown, E. (2006). Written feedback for students: too much, too detailed or too incomprehensible to be effective? *Bioscience Education, 7*(1), 1-16. doi:10.3108/beej.2006.07000004

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. doi: 10.3102/003465430298487

Karpicke, J. D., & Aue, W. R. (2015). The Testing Effect Is Alive and Well with Complex Materials. *Educational Psychology Review, 27*(2), 317-326. doi:10.1007/s10648-015-9309-3

Karpicke, J. D., & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science, 319*(5865), 966-968. doi: 10.1126/science.1152408

Lucassen, M. (2018). *Taxonomie van Bloom* [Online afbeelding]. Geraadpleegd op 14 november 2019, van https://www.vernieuwenderwijs.nl/de-taxonomie-van-bloom-vaak-verkeerd-gebruikt-maar-zo-werkt-het-wel/

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., Jaeger, A. (2019). Retrieval Practice in Classroom Settings: A Review of Applied Research. *Frontiers in Education, 4.* doi:10.3389/feduc.2019.00005

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218. doi: 10.1080/03075070600572090

Nitko, A. J., & Brookhart, S. M. (2007). Educational assessment of students (5th ed.). Upper Saddle River, NJ: Pearson Education.

Onderwijsraad. (2018). *Toetswijzer: Naar een eigen(tijdse) wijze van toetsen en examineren.* Geraadpleegd op 21 oktober 2019, van https://www.onderwijsraad.nl/publicaties/adviezen/2018/12/13/toets-wijzer

Pan, S. C., & Agarwal, P. K. (2018). *Retrieval Practice and Transfer of learning. Fostering students' application of knowledge.* UCS: San Diego.

Pyc, M. A., & Rawson, K. A. (2011). Costs and benefits of dropout schedules of test-restudy practice: Implications for student learning. *Applied Cognitive Psychology, 25*(1), 87-95. doi: 10.1002/acp.1646

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends In Cognitive Sciences, 15*(1), 20-27. doi:10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1(3),* 181-210. doi:10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242-248. doi:10.1016/j.jarmac.2012.09.002

Rowland, C.A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing-effect. *Psychological Bulletin, 140*(6). doi:10.1037/a0037559

Schildkamp, K., Heitink, M., Van der Kleij, F., Hoogland, I., Dijkstra, A., Kippers, W., & Veldkamp, B. (2014). *Voorwaarden voor effectieve formatieve toetsing: Een praktische review.* Geraadpleegd op 16 oktober 2019, van https://www.nro.nl/wp-content/uploads/2015/01/Eindrapport-formatief-toetsen-Utwente-revisie.pdf

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago, IL: Rand McNally.

Sluijsmans, D., Joosten-Ten Brinke, D., & Van Der Vleuten, C. (2013). *Toetsen met leerwaarde: Een reviewstudie naar de effectieve kenmerken van formatief toetsen*. Geraadpleegd op 14 oktober 2019, van https://www.nro.nl/wp-content/uploads/2014/05/PROO+Toetsen+met+leerwaarde+Dominique+Sluijsmans+ea.pdf

Stenlund, T., Sundström, A., & Jonsson, B. (2014). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology, 36*(10), 1710-1727. doi: 10.1080/01443410.2014.953037

Van Berkel, H., Bax, A., & Joosten-ten Brinke, D. (2017). *Toetsen in het hoger onderwijs* (4e editie).

Houten, Nederland: Bohn Stafleu van Loghum.

Van Gog, T., & Sweller, J. (2015). Not New, but Nearly Forgotten: The Testing Effect Decreases or even Disappears as the Complexity of Learning Materials Increases. *Educational Psychology Review*, *27*(2), 247–264. doi:10.1007/s10648-015-9310-x

Yeo, D. J., & Fazio, L.K. (2019). The optimal learning strategy depends on learning goals and processes: Retrieval practice versus worked examples. *Journal of Educational Psychology*, *111*(1), 73-90. doi: 10.1037/edu0000268

Bijlage

Bijlage A: Oefentoets (met toelichtende feedback)

Anatomie, fysiologie en pathologie

1. Colloïdalen spelen meer een rol bij de bloedstolling dan bij de osmotische zuigkracht.
   ONJUIST
   Feedback:
   Eiwitten hebben een osmotische waarde aan de veneuze kant van capillairnetwerk. Trombocyten spelen een rol bij de stolling.

2. De eiwitten in onze voeding zijn eerder bedoeld voor de opbouw van spieren, dan voor het leveren van energie.
   JUIST
   Feedback:
   Glucose, suikers en koolhydraten zijn bedoeld voor het leveren van energie. Eiwitten zijn behulpzaam bij de opbouw van spieren.

3. Een groot regeneratievermogen is meer een kenmerk van epitheelweefsel dan van kraakbeen.
   JUIST
   Feedback:
   Het regenereren (het opnieuw maken) is bij uitstek epitheelweefsel. Dit is dekweefsel dat alle oppervlakken van het lichaam bedekt. Een voorbeeld hiervan is wondweefsel.

4. In rust pompt het hart van een volwassene per 24 uur eerder 2000 liter bloed dan 8000 liter bloed rond.
   ONJUIST
   Feedback:
   Per minuut wordt 5 liter bloed rondgepompt. Dat betekent 300 liter bloed per uur. Vervolgens wordt deze 300 liter vermenigvuldigd met 24 (uur). Dan wordt er ongeveer 7200 liter per 24 uur rondgepompt. Dit komt dichterbij de 8000 dan bij de 2000 liter.

5. Medicijnen kennen een 'first-pass effect'. Dit betreft meer de eerste passage door de nieren dan door de lever.
   ONJUIST

   Feedback:

Het 'first-pass effect' wordt opgenomen via de bloedvaten in de darmen. Vanuit de darmen gaat het naar de poortader en vervolgens naar de lever. De nieren spelen hierbij dus geen rol.

6. De meest bedreigde organen bij een shocktoestand zijn eerder 'het hart en de nieren' dan de 'milt en de lever'.
   JUIST
   Feedback:
   Alhoewel de milt en de lever belangrijke organen zijn, betreffen het niet de vitale organen zoals het hart en de nieren. Relateer dit aan de ABCDE-methode. Hierbij worden niet de milt en de lever beoordeeld.

7. Bij een gezond persoon is het aantal trombocyten in het bloed groter dan het aantal erytrocyten in het bloed.
   ONJUIST
   Feedback:
   Erytrocyten (rode bloedcellen) zijn verantwoordelijk voor het bloedtransport. Trombocyten (bloedplaatjes, kleinste cellen) zijn belangrijk voor de bloedstolling na verwonding. De gehele dag heeft je lichaam zuurstof nodig, dus is het aantal erytrocyten groter dan de trombocyten. Een lichaam heeft niet de gehele dag stolling nodig.

8. De eerste fase in wondgenezing is eerder gericht op bloedstelping dan op regeneratie van het weefsel.
   JUIST
   Feedback:
   In eerste instantie worden alle bloedplaatjes naar de wond gedirigeerd/verplaatst om de bloeding te stoppen.

9. De kleine bloedcirculatie voorziet het longweefsel van zuurstof.
   ONJUIST
   Feedback:
   De grote bloedsomloop (zuurstofrijk) voorziet de cellen en weefsel van zuurstof. De functie van de kleine bloedsomloop is dat de rechterkant van het hart zuurstofarm bloed naar de longen verpompt om daar de longen van zuurstof te voorzien. Longweefsel wordt voorzien van zuurstof door de aftakking van de grote circulatie.

10. Skeletspieren zijn eerder opgebouwd uit glad spierweefsel dan uit dwarsgestreept spierweefsel.
    ONJUIST

Feedback:
Glad spierweefsel functioneert zelfstandig in tegenstelling tot dwarsgestreept spierweefsel. Skeletspieren functioneren niet zelfstandig dus is er sprake van dwarsgestreept spierweefsel. Een voorbeeld van gladspierweefsel (dat zelfstandig functioneert) is het hartspierweefsel.

Kern

11. Vermoeidheid bij decompensatio cordis (hartfalen) valt eerder onder de dimensie 'mentaal welbevinden' dan onder de dimensie 'lichaamsfuncties'
    ONJUIST
    Feedback:
    Binnen het model van Positieve Gezondheid vallen slaap en lichamelijke condities onder lichaamsfuncties.

12. De gezondheidsdeterminanten zijn onderverdeeld in drie soorten beïnvloedende factoren. Eén van deze beïnvloedende factoren is eerder het 'mentaal functioneren' dan 'leefstijl'.
    ONJUIST
    Feedback:
    De gezondheidsfactoren zijn opgedeeld in drie soorten beïnvloedende factoren, namelijk: omgevingsgebonden factoren, persoonsgebonden factoren en leefstijlgebonden factoren. Mentaal functioneren behoort daarom niet tot een van de drie beïnvloedende factoren.

13. Bij Evidence Based Practice wordt er onderscheid gemaakt tussen drie kennisbronnen. Een van deze kennisbronnen is eerder 'expert opinies' dan het 'verpleegkundig proces'. JUIST
    Feedback:
    Het model van Evidence Based Practice bestaat uit: opinie, wetenschappelijk onderzoek en patiëntervaringen.

14. De gezondheidspatronen van Gordon worden eerder gebruikt in de anamnesefase dan in de diagnosefase.
    JUIST
    Feedback:
    De 11 gezondheidspatronen van Gordon zorgen ervoor dat de hulpverlener inzicht krijgt in de cliënt. Deze kennis, vergaard tijdens de anamnesefase, draagt bij aan het stellen van de juiste diagnose en het behandelplan.

15. De hielprik is eerder een vorm van primaire preventie dan secundaire preventie.
    ONJUIST

Feedback:
De hielprik valt onder de secundaire preventie. Vroegtijdige signalering betreft secundaire preventie. Voorkomen behoort bij primaire preventie.

16. Bij ambivalente hechting vertoont een kind eerder wisselend gedrag dan vermijdend gedrag.
JUIST
Feedback:
Ambivalent betekent wisselend of tegenstrijdig. Bij ambivalente hechting vertoont het kind dus afwisselend gedrag waarbij het opzoeken van nabijheid en afwerend gedrag afwisselen met boosheid en verdriet.

17. Je vermoedt dat een kind bij jou in de straat mishandeld wordt.
Volgens de Meldcode kindermishandeling ben je verplicht om hiervan een melding te maken.
ONJUIST
Feedback:
De meldcode is nadrukkelijk geen meldplicht. Organisaties zijn verplicht een meldcode te hebben, professionals zijn verplicht te handelen volgens de stappen van de meldcode, maar het doorlopen van deze meldcode hoeft niet te leiden tot een melding. In de Wet meldcode is wel opgenomen dat professionals het recht hebben te melden bij Veilig Thuis. Ook als gezinsleden daar geen toestemming voor geven. Het meldrecht houdt in dat professionals persoonsgegevens van volwassenen en kinderen mogen doorgeven aan Veilig Thuis, zodat zij een onderzoek naar de gezinssituatie kan starten. Bovendien mogen professionals informatie geven als Veilig Thuis daar vanwege haar onderzoek om vraagt.

18. In het verpleegkundig proces wordt de diagnosefase eerder gevolgd door de interventies dan door het vaststellen van de gewenste resultaten.
ONJUIST
Feedback:
Het verpleegkundig proces heeft in theorie een vaste volgorde. Na de diagnosefase worden eerst de gewenste resultaten vastgelegd om vervolgens interventies te plannen.

19. ADD en ADHD zijn beide ontwikkelingsstoornissen.
Het kenmerk 'snel afgeleid zijn' is eerder een kenmerk van ADHD dan van ADD.
ONJUIST
Feedback:

De bekendste variant is natuurlijk ADHD zelf (dus mét de H), en dan is het kind vooral erg druk. Als het kind de tweede variant heeft (ADD) dan is het kind voornamelijk snel afgeleid. De derde variant is ADHD van het gecombineerde type. Dan is het kind én heel druk én snel afgeleid.

20. Dhr. van Stiphout is opgenomen op de afdeling Cardiologie van het Maxima Medisch Centrum vanwege decompensatio cordis (hartfalen). Verpleegkundige Wouter gaat met dhr. in gesprek over zijn lichamelijke klachten zoals vermoeidheid en kortademigheid en het effect hiervan op het sociale leven, de mentale gezondheid en het dagelijks functioneren.
De benadering van Wouter is eerder te omschrijven als holistisch dan als dualistisch.
JUIST
Feedback:
De holistische benadering gaat uit van de mens als geheel (lichamelijk, mentaal en sociaal). Een dualistische benadering maakt een strikt onderscheid tussen lichaam en geest.